Chapter 41
Deep Dive Speech Technology

Marcin Skowron, Gerhard Backfried, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernáez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, and Petr Schwarz

Abstract This chapter provides an in-depth account of current research activities and applications in the field of Speech Technology (ST). It discusses technical, scientific, commercial and societal aspects in various ST sub-fields and relates ST to the wider areas of Natural Language Processing and Artificial Intelligence. Furthermore, it outlines breakthroughs needed, main technology visions and provides an outlook towards 2030 as well as a broad view of how ST may fit into and contribute to a wider vision of Deep Natural Language Understanding and Digital Language Equality in Europe. The chapter integrates the views of several companies and institutions involved in research and commercial application of ST.[1]

Marcin Skowron · Gerhard Backfried
HENSOLDT Analytics GmbH, Austria, marcin.skowron@hensoldt.net, gerhard.backfried@hensoldt.net

Marek Kováč · Johan Rohdin · Petr Schwarz
Phonexia, Czech Republic, kovac@phonexia.com, rohdin@phonexia.com, schwarz@phonexia.com

Eva Navas · Inma Hernáez · Jon Sanchez · Ibon Saratxaga
University of the Basque Country, Spain, eva.navas@ehu.eus, inma.hernaez@ehu.eus, jon.sanchez@ehu.eus, ibon.saratxaga@ehu.eus

Aivars Bērziņš
Tilde, Latvia, aivars.berzins@tilde.com

Joachim Van den Bogaert
CROSSLANG, Belgium, joachim.van.den.bogaert@crosslang.com

Franciska de Jong
CLARIN ERIC, The Netherlands, franciska@clarin.eu

Andrea DeMarco · Michael Rosner
University of Malta, Malta, andrea.demarco@um.edu.mt, mike.rosner@um.edu.mt

Peter Polák
Charles University, Czech Republic, polak@ufal.mff.cuni.cz

[1] This chapter is an abridged version of Backfried et al. (2022).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_41

1 Introduction

Speech – as the most natural manner for humans to interact with computers – has always attracted enormous interest. Speech Technology (ST) has been a focus of research and commercial activities over the past decades. From humble beginnings in the 1950s, they have come a long way to current state-of-the-art approaches.

Stimulated by a shift towards statistical methods, the 1980s witnessed an era of Hidden-Markov-Models (HMM), Gaussian-Mixture-Models (GMM) and word-based *n*-gram models combined into speech recognition engines employing ever more refined data structures and search algorithms (Jelinek 1998). The availability of data to train these systems was limited to only a few languages, often driven by security and commercial interest. Even then, work on neural networks (NN) was already being carried out and viewed by many as the most promising approach. However, it was not until later (2000s) that the availability of training data paired with advances in algorithms and computing power finally began to unleash the full potential of NN-based ST. Especially over the past couple of decades, ST has evolved dramatically and become omnipresent in many areas of human-machine interaction. Embedded into the wider fields of Artificial Intelligence (AI) and Natural Language Processing (NLP), the expansion and scope of ST and its applications have accelerated further and gained considerable momentum. Recently, these trends were complemented by a paradigm shift related to the rise of language models (Bommasani et al. 2021), such as BERT (Devlin et al. 2019) or GPT-3 (Brown et al. 2020): models trained on a broad scale, adaptable via fine-tuning and able to perform very well on a wide range of tasks. Substantial advances in algorithms and high-performance hardware have led to massively increased adoption and further technological improvements. With speech and natural language forming fundamental pillars of human communication, ST may now even be perceived as "speech-centric AI".

With the emergence of intelligent assistants, ST has become ubiquitous, yet many ST systems can only cope with restricted domains and can be used only with the most widely spoken languages. For languages with a low number of speakers, ST systems are still all but absent or severely limited in their scope. Recent advances in Machine Learning (ML) and ST have begun to enable the creation of models also for such less well-resourced languages. However, these approaches are generally more complex, expensive and less suitable for wide adoption. While recently presented results indicate that novel approaches could indeed be applied to address some of the challenges related to low-resourced languages, the scope of their application and inherent limitations are still the subject of ongoing research (Lai et al. 2021).

STs have been investigated and researched in their own right. However, their full potential often only becomes evident when combined with further technologies forming intelligent systems capable of complex interaction, encompassing a diverse set of contexts and spanning multiple modalities. To the casual user, individual components then become blurred and almost invisible with one overall application acting as the partner within an activity which may otherwise be carried out together with a fellow human being. In this setting, the aggregation of technologies goes beyond narrow and highly specialised systems towards combined and complex systems, pro-

viding a notion of a more general and broader kind of intelligence. Speech and language, as the most natural vehicles for humans to communicate with machines, thus become the gatekeepers to and core of a broader kind of AI.

1.1 Scope of this Deep Dive

The scope of this deep dive encompasses a wide range of STs including language identification, speaker recognition, automatic speech recognition, technologies addressing paralinguistic phenomena as well as text-to-speech. It gathers and synthesises the perspectives of European research and industry stakeholders on the current state of affairs, identifies several main gaps affecting the field, outlines a number of breakthroughs required and presents the technological vision and development goals for the next years. In line with the other deep dives in this book, we adopt a multidimensional approach where both market/commercial as well as research perspectives are considered and concentrate on the following aspects: technologies, models, data, applications and the impact of ST on society. The tendency for the combination of technologies into more powerful systems, encompassing several individual technologies and models, has become apparent and is reflected throughout this chapter.

1.2 Main Components

STs encompass technologies on the recognition as well as production side of speech. They comprise a wide spectrum of sub-fields such as automatic speech recognition (ASR), the identification of language or dialects, speaker recognition/identification (SR/SID), the detection of age and gender, emotions, paralinguistic traits and the production of synthesised speech (often called text-to-speech).

2 State-of-the-Art and Main Gaps

2.1 State-of-the-Art

Traditional ASR systems consist of components for audio pre-processing, an acoustic model, a pronunciation model as well as a language model defined over units of a lexicon. Within a search algorithm, these elements are combined to produce the most likely transcript given the input audio. In this scheme, models generally are of a generative nature and optimised individually. Since the early 2000s, these components are being replaced with deep neural networks (DNNs). This change was made possible by advances in algorithms and models as well as the massive increase in available training data and computing power (GPUs). As a result, word error rates

(WERs) could be reduced considerably in many domains and languages. However, the performance of ASR systems still varies dramatically depending on the domain and language, with low-resource languages still exhibiting WERs resembling those of English many years ago.

For applications in practice ("ASR in the wild"), hybrid systems combining elements such as HMMs and DNNs still dominate the state of play. As such, they can still be regarded as state-of-the-art outside of research labs. Toolkits like Kaldi provide a sound basis for the development of systems for research as well as commercial environments. Novel approaches in the area of self-supervised learning, e. g., Wav2Vec 2.0 by Facebook (Baevski et al. 2020), focus on leveraging vast amounts of unlabelled data. Latent representations of audio are produced representing speech sounds similar to (sub-)phonemes which are then fed into a Transformer network. This approach has been shown to outperform other typical paths of semi-supervised methods, while also being conceptually simpler to implement and execute. The possibility to employ smaller amounts of labelled data as well as being able to train multilingual models provide strong arguments for such approaches.

Typically, ASR outputs unstructured and normalised text without punctuation marks. This is not problematic in use-cases where the user input is short and concise, e. g., when asking a question to a virtual assistant. However, when generating transcripts for longer speech, it is crucial to restore punctuation to improve readability and provide structure to the transcript. Moreover, punctuation is relevant for further downstream tasks such as named-entity recognition (NER), part-of-speech (POS) tagging and machine translation (MT). Recognition errors introduced by ASR may lead to cascaded errors in these tasks, e. g., for MT (Ruiz et al. 2019).

State-of-the-art SR systems use neural networks to extract a representation (embedding) for the speaker in an utterance. The input to the network typically consists of features extracted from frames of 20-30ms, although there are also ongoing efforts to take the raw waveform as an input. Embeddings are then compared in order to decide whether they are from the same person or not. Typical NN architectures for embedding extraction are TDNN, ResNet, or LSTM. The standard choice of backend is a generative model: Probabilistic Linear Discriminant Analysis (PLDA). Recently, using cosine similarity plus an affine transform has proven to yield competitive performance. An advantage of generative backends is that scoring with different numbers of enrolment utterances becomes trivial. In addition to variations of the embedding extractor architecture, many recent research efforts have focused on the training objective. If the task at hand is verification, the most intuitive manner would be to train the extractor for this task. However, in practice, it often works better to train the extractor for classification. That is, for a training utterance the network should classify who among the speakers in the training set speaks in the utterance.

State-of-the-art language identification (LID) systems are based on DNNs ingesting sequences of frame-level features as input, processing them and applying a pooling mechanism to obtain an utterance level representation which is eventually classified. During training, this whole chain is performed in an end-to-end (E2E) fashion. In testing, either the trained DNN is used directly for classification or the utterance

level representations can be extracted and used in a simple backend for classification, e. g., a Gaussian linear classifier.

In the field of Speech Emotion Recognition (SER), a wide range of methods have been used to extract emotions from signals. Similar to other ST domains, Deep Learning is rapidly becoming the method of choice and several E2E models have been proposed (Tang et al. 2018). Unlike ASR, these have not yet become part of our everyday lives. To achieve this goal, SER systems require more accurately labelled data to improve training accuracy, more powerful hardware to speed up processing, and more powerful algorithms to improve recognition rates. In addition, further insights from fields such as psychology or neurology may be required. Detecting the cognitive states and reactions of a user is a step towards designing proactive systems capable of adapting to the user's needs, preferences and abilities. As in other related ST-fields, the detection of personality traits, mood disorders, signs of depression and other medical conditions has found its application in recent years. Techniques based on automatic processing of the voice signal have been used for language and cognitive assessments. These approaches provide the means for quantifying signal properties relevant for the detection of specific pathologies. Due to the development of automatic methods facilitating the evolving control of a wide population suffering from Alzheimer's disease, a number of industry applications aimed at the detection of neurodegenerative disorders have been introduced.

Neural networks have greatly impacted the speech synthesis field by improving the quality and naturalness of synthetic voices compared to traditional systems and by enabling training in an E2E fashion. While traditional multi-stage pipelines are complex and require extensive domain expertise, E2E systems reduce the complexity by extracting the audio directly from the input text without requiring separate models. E2E text-to-speech (TTS) systems have shown excellent results in terms of audio quality and naturalness. However, they usually suffer from low training efficiency, requiring large sets for training. Full E2E architectures have been proposed, e. g., FastSpeech 2 (Ren et al. 2021). These systems produce spectrograms from text by applying an encoder-decoder architecture that produces a latent representation of the input text (or phonetic transcription) which is subsequently transformed into spectrograms. These systems provide outstanding results in terms of the quality and naturalness of the generated voices but require large amounts of high-quality recordings to be trained properly. Efforts are being made to deploy these systems for low-resource languages by improving data efficiency, applying transfer learning or training multilingual models. Other areas of intense research activity are style transfer, controllable and expressive voice generation, new efficient neural vocoders and speaker adaptation with a reduced amount of data. Regarding expressive speech synthesis, Global Style Tokens (Wang et al. 2018) represent one of the most common approaches. It consists of a reference encoder, encoding the speech Mel-spectrogram, and a style token layer, learning different prosodic aspects in a set of trainable embeddings. The reference embedding is compared with each style token with the help of a sequence-to-sequence multi-head attention module, forming a weighted sum of the style tokens called "style embedding". This style embedding is then concatenated to the text encoder output, thus conditioning the Mel-spectrogram

synthesis on both text and encoded prosody of the speech. Other popular methods include Flowtron (Valle et al. 2021), Mellotron (Valle et al. 2020), and Ctrl-P. Developing high-quality synthetic voices with DNN-based techniques requires large amounts of high-quality recordings from a single speaker. This requirement is often difficult to fulfil, especially for minority languages and dialectal speech. The generation of new synthetic voices is also hindered by this extensive data requirement. Efforts are being made to share data among languages and speakers in order to train the common aspects more robustly. Multi-speaker and multi-language modelling is a common strategy in DNN-based TTS synthesis to achieve improved voice quality with a reduced amount of data from a single speaker. However, the quality of these voices is not yet comparable to those obtained with large databases.

2.2 Main Gaps

While ST has found its way into a series of application fields, various important issues have not been addressed thoroughly and remain active areas of research. In the following, we review the main gaps and present them in the context of global and regional business activities, requirements related to the availability of qualified personnel, privacy and trust concerns, as well as technical and end-user perspectives.

Effects of scale – A trend towards increasingly complex E2E systems can be observed in all areas of ST. Due to the extreme demand on resources, e. g., data, compute, energy, or infrastructure, the construction of such models is limited to a handful of actors. The activities to make pre-trained language models available for transfer learning and fine-tuning and to allow others to also participate in major advances are certainly beneficial. However, the extent of this transfer and level of control in the hands of a few institutions poses a risk to other actors, to the market and potentially even to innovation in the sector as a whole. Compared to the US and China, European players are at a stark disadvantage concerning resources, i. e., data, technology and funding. Academic institutions risk lagging behind industrial research due to a lack of resources and may have to rely on national initiatives to keep up.

Trained personnel and expertise – A further gap, concerning all areas of speech processing, can be identified in the scarcity of trained personnel and expertise as well as the risk of losing emerging talent to innovative power-players outside of Europe (with possibilities and employment conditions which generally cannot be matched by European players). Even in light of the democratisation of technology and auto-ML, allowing a much broader audience to create models and deploy these for use, respective educational programmes in speech (and NLP/LT) technologies form the foundation for future European success in these areas and may hinder it if not appropriately established and strengthened.

Privacy and trust – Data leaks and scandals in recent years have spurred the interest of individuals as well as of policy-makers. Concerns have arisen regarding trust, privacy, intrusion, eavesdropping, or the hidden collection and use of data. These

concerns have been recognised by many actors but are only addressed to a very limited extent, as they often counteract commercial interests.

Technical perspectives – The focus of many ST fields on rather constrained conditions has left gaps in more diverse settings such as: processing of distant speech; noisy environments; accented speech, non-native speech, dialectal speech, code-switching, spontaneous, unplanned speech, emotional speech and connected aspects concerning sentiments expressed; the integration of ST into collaborative environments, multiple, simultaneous speakers engaged in vivid discussions; as well as the integration of paralinguistic aspects and technologies.

Group settings, multiple-user scenarios – While most research focuses on a single user's interactions, STs embodied in virtual assistants are becoming increasingly popular in social spaces. This highlights a gap in our understanding of the opportunities and constraints unique to multiple user scenarios. These include detecting whether users are addressing the system or other participants, speaker diarisation, aspects of social dynamics, and finding interaction barriers. Due to these factors, the usefulness of voice interfaces in group settings is still restricted.

Interdisciplinary research work (Digital Humanities and Social Sciences and the Humanities, SSH) – While the connection to the field of digital humanities and computational social sciences is not firmly established yet, it could be beneficial to set up collaborative links with a range of disciplines and domains working with spoken data. In particular, the insights and requirements stemming from the needs for transcription workflows and audio mining tools of communities producing and (re)using oral history data and interview recordings may help identify gaps in language resources for model training and domain adaptation (Draxler et al. 2020). It could be beneficial to identify imbalances in language-specific support for the recognition, annotation and retrieval of the types of structured conversational speech that are used in interview settings in SSH and beyond (Pessanha and Salah 2022).

Challenges related to an increased modelling power – The increase in modelling power and performance achieved over the last years also comes with some drawbacks and challenges. These include a need for even more data, respectively a lack of interest and work on the creation of new paradigms using less data. Current approaches include shallow and deep fusion, but the question of how to optimally combine language models (LMs) and DNN structures has still not been addressed comprehensively. Models requiring the complete input sequence for processing do not match well with requirements to perform causal processing. Several attempts to enable causal processing are being explored, among them the use of neural transducers running processing at regular intervals. The extent of context may also incur additional processing costs which need to be balanced and mitigated.

Models: interoperability and transparency – Models are not transparent and thus hard to interpret. This is partly due to the fact that previously individual components have been combined into single models. The complex process of hyper-parameter tuning is often too resource-intensive and thus has not been addressed in many instances. Elements of input/output like byte-pair-encodings (BPE) have been suggested but these contradict the idea of genuine E2E processing. Integration of several components into one model prompts the question of whether further downstream

technologies will also become part of such integrated models. The combination in turn raises questions about the interpretability and transparency of such systems.

Explainability and transparency for critical methods and technologies – While in the last decade, ST research has achieved improvements in terms of performance, progress in terms of understanding of the architectures used and of the nature of the data and task has been limited. This is partly due to the fact that the NNs used in modern systems are harder to understand than the generative models of previous generation systems. It is also due to a lack of interest from the industry and funding agencies to support this type of research. Students are also generally inclined to work on topics that mainly aim at improving performance since this increases their chances of obtaining a well-paid job in the industry after graduation.

End-users' perspective – STs have made a leap in becoming adopted in many settings for commercially attractive languages. Especially the proliferation of intelligent Voice Assistants (VAs) has made speech a common mode of interaction. However, several issues limiting the further adoption and widespread use of ST remain: these include problems in accurately recognising accented speech, a lack of trust in VAs to execute more complex or sensitive tasks, and concerns related to privacy and data collection. This issue is further exacerbated by the fact that systems often operate in the cloud rather than on-premise. Many VAs may already be utilised in languages other than English, but coverage and supported functionality vary greatly. The gaps in support create barriers for users whose primary language is not fully catered for, or supported only to a limited extent, forcing them to communicate in a non-native language or risk being excluded from using the ever more popular systems and services. This way, non-native users are pushed to develop different strategies and modes of interaction, including a reduced level of language production and more frequent use of visual feedback.

Data: availability, diversity – The main challenge related to data concerns its availability, i. e., adequate datasets for low-resource languages of an appropriate amount and quality. Various efforts aim to mitigate this fact by focusing on transfer learning and fine-tuning of models. However, whereas this approach is certainly beneficial, it generally does not yield models of equal performance as for languages equipped with large amounts of training data. The lack of data for low-resource languages effectively excludes certain approaches from being applied.

Data: diversity of voices – Some public databases available to train DNN-based TTS systems are only useful for building monolingual neutral voices for a number of major languages. The availability of open data free of restrictions such as copyright and limitations due to GDPR regulations in the remaining major languages and all minority languages would allow the development of TTS systems for these languages too. Databases with more expressive and spontaneous recordings are needed to build TTS systems suitable for more emotion-demanding applications like audiobook reading, movie dubbing and HCI. The vast majority of datasets correspond to adult voices and there is a lack of data to generate child and elderly voices. As the voice is an important component of our identity, more diverse datasets are needed to generate personalised voices that can suit any user.

Accuracy: reaching usable thresholds for applications – The single most frequently mentioned hindering factor for the broad adoption of ST is one that has been mentioned for the past 40 years, namely accuracy. The perceived accuracy and its exact meaning have changed dramatically: from individual words being misrecognised to intentions that are not correctly interpreted in complex situations. For example, WER as an evaluation measure has had its merits in measuring progress in ASR (and still does). However, more comprehensive approaches to measuring the impact of ASR performance on downstream tasks and actual deployments may require novel measures. WER alone clearly does not provide the full picture when it comes to the perceived performance and usability of complete systems comprising several kinds of STs and LTs. Regarding TTS, accuracy translates to a lack of naturalness and robustness of the synthesised speech. Different approaches have been taken, some of them focused on designing robust attention mechanisms, others including alignment information at the input, or substituting the attention mechanism with networks that can predict the estimated duration of the input phonemes. However, the problem has not been solved completely yet and keeps hindering the practical application of TTS systems in many instances. For SR, technologies have already reached acceptable performance for many applications. However, this does not mean that there is no need or opportunity for further research. All applications of SR would benefit from better core performance and increased robustness to different acoustic conditions and other variables occurring in real-world speech data.

Dialectal speech and multilingual training – Most ST systems process speech only in the main variety of languages. To date, little attention has been devoted to dialectal speech. Certain STs can be used in languages different from the one(s) they were originally designed for. However, the performance of such systems typically deteriorates. Some progress has been made to make systems more language-independent (e. g., multilingual training, adversarial adaptation), but there is still ample room for improvement. The effectiveness of such approaches for languages that differ substantially from those used in training has not been investigated thoroughly and warrants further work.

3 The Future of the Area

3.1 Contribution to Digital Language Equality

Purely technological systems alone do not exist – they are always embedded in a social context and should thus always be viewed as socio-technical systems. The applications of ST have diverse and multifaceted impacts on several key aspects for societies. Technologies reaching performance levels resembling those of humans may in many aspects lead to a humanisation of technology, ascribing human attributes to system behaviour. Patterns of human-to-human (H2H) interaction may be applied to human-to-machine (H2M) interaction leading to heightened expectations and potentially to subsequent disillusion.

Digital language inequality – The unbalanced availability and quality of ST resources strongly impact the performance of systems for different groups of languages. For languages supported to a lesser extent, performance and accuracy are typically significantly lower compared to resource-rich languages. In extreme cases, selected functionalities or support for such languages may not be available at all. In addition, language varieties, dialects or accents may not be supported or only supported on very limited levels. STs are thus not accessible nor available to everyone on an equal level. The lack of commercial interest in the long tail of "small languages" translates to a significantly slower pace of ST improvements and commercial adoption for the latter group. For native speakers of these languages, these imbalances lead to wider usage of the better-supported major languages, such as English. Motivating speakers to use these major languages more frequently creates a new set of challenges related to handling accented and non-native speech. Compared to the level of service and the support provided for native speakers, this results in lower performance, weakened experience and reduced usability, rendering ST less useful or even useless in the extreme case.

Energy consumption and sustainability – The growing energy consumption required for the ever-expanding amount of data being processed and the tendency towards continuously more complex ST models have become evident since the race for the largest models has been going on. Due to the extreme demand on resources, the generic construction of complex AI, NLP and ST systems is typically limited to a few actors. Surging interest in sustainability may cause actors to reconsider the massive increase in energy consumption that currently often accompanies progress in ST. An opportunity (and marketing advantage) may arise from directing efforts towards the creation of high-performance/low energy-consumption ST, exploring the capacities of E2E or novel direct speech-to-speech systems to lower the energy consumption by avoiding a separate, cascading training of sub-systems.

Labour market – A further economic aspect concerns the impact of ST on automation and as a consequence on the job market as a whole. As technologies such as chatbots are being adopted in pursuit of efficiency, they also perform an increasing number of tasks previously reserved for humans. ST and AI thus blur the boundary between humans and technology leading to shifts in jobs and even entire industries. Clearly, a message of cooperation and support rather than of rivalry and replacement needs to be communicated and acted upon.

Politics and democracy – It has been pointed out that language strongly influences the manner in which we think and argue about political issues. Language causes mental frames to be activated and form our portfolio of ideas. Politicians and influencers have long discovered these mechanisms and are applying them actively to push their respective agendas. Having this central and immediate effect on cognitive mechanisms, linguistic plurality also forms the basis of cognitive plurality and as such plays a fundamental role in securing diverse and democratic values. Limitation to a few individual languages – such as may happen due to limited digital support for certain languages – impoverishes and reduces this variety, the flexibility and spectrum for expression of thoughts and (political) ideas.

Biases and ethical issues – Several ST systems have been shown to be less accurate for female speakers than for males. This is not because women are underrepresented in the training data but more likely due to the properties of female and male voices. Various ethnic groups may be underrepresented in datasets and consequently, performance becomes less accurate. It should also be noted here that being in a group for which a system performs worse can be either an advantage or a disadvantage depending on the application and the type of error the system tends to commit more often (false positives or false negatives). Another ethical concern pertaining to ST is due to possible privacy breaches through mass surveillance. TTS systems have reached a quality level and degree of similarity with the voice of humans that could be used to generate deep-fake voices or voices of deceased persons. Despite this scope for misuse, most of the possible applications of high-quality voices are positive, and people with speech disorders, visual impairment and other disabilities could greatly benefit from them. However, deep-fakes could also be employed for illegal activities such as committing fraud or discrediting people. New regulations and the development of ad hoc legislation are critical to mitigating this pernicious effect. Tools able to detect speech deep-fakes need to be produced, and anti-spoofing techniques that discriminate synthesised from natural speech must be developed in close collaboration with teams working in ST.

Users with special needs – While ASR systems achieve great accuracy on standard speech, they perform poorly on disordered speech and other atypical speech patterns. Personalisation of ASR models, a commonly applied solution to this problem, is usually performed on servers posing problems related to data privacy and data transfer. While on-device personalisation of ASR has recently shown promising results in a home automation domain for users with disordered speech (Tomanek et al. 2021), more research is required to increase performance for these groups of users and provide support for open conversations. TTS is considered an assistive technology and as such, it may contribute to the integration of individuals with visual impairments or learning disabilities. By developing robust TTS systems, these people could enjoy the same advantages as any person without a disability. It also facilitates equal access to education and supports foreigners who may struggle with the language. ST can contribute to the integration of immigrants by making it easier to learn local languages and can help people with literacy issues and pre-literate children to access content presented in written form. ST may also prove helpful in times of aging populations with degrading eyesight. Integrated into virtual assistants, STs are able to provide support to elderly people, assisting them with reminders of appointments and medication needs, providing access to online information and improving both their ability to live by themselves and strengthen their autonomy. Another particular benefit of TTS relates to orally impaired people. Voice is an essential component of our identity that we usually take for granted. However, losing it can affect how others perceive us and our own sense of who we are. TTS technology is able to provide a voice for those who have lost their own via personalisation suiting the characteristics desired by each user.

Privacy and trust – As technologies are entering the homes and offices of users on a broad scale, an enhanced level of attention to privacy concerns, ethics and policy

is essential. Policymakers, policy watchdogs, the media and consumers alike need to assume the role of gatekeepers. Trust is viewed as the main currency and key to the adoption and acceptance of technologies. Scandals and opaque behaviour on the part of ST providers may have detrimental effects. Whenever ST is linked to a person's identity and used for access control or authorisation, the issue of trust becomes especially important. For example, STs are used to authorise access to resources such as a bank account or building. In surveillance applications, it is used for detecting and identifying criminals. In forensics, SR is used for comparing a voice recording from a crime scene with the voice of a suspect or a victim. For voice assistants, SR can be essential to make sure that certain requests are fulfilled only if made by the owner of the respective device or commodity. All of the above applications rely on high-performance and trusted ST, and can benefit tremendously in commercial terms if applied within these contexts. Many applications of ST store audio in the cloud. It is essential to secure guarantees regarding how data is used or will be used in the future by cloud service providers (the risk of leaking always remains). In the long run, the question will be whether any possible breaches, leaks or scandals involving ST will erode trust to a level that users will no longer volunteer to provide their data. Of course, the distrust will be weighed against the commodity of using certain devices and platforms whose terms of use may simply require the user to do so. Opting out may not always be a realistic option.

Unlawful surveillance – A further area of concern is the extent of unlawful surveillance by governments, state agencies or corporations, infringing citizens' rights, liberties, adversely affecting public discourse, democratic values and influencing the political powers (Stahl 2016). The concerns comprise privacy invasion, accountability of intelligence and security services, and the (non-)conformity of mass surveillance activities with fundamental rights (Garrido 2021). Their effects on the social fabric of nations can only be considered and analysed jointly with the rapidly extending technological capacities and the pervasiveness of devices able to capture, process and transmit relevant data. Regardless of the form of government, the growing extent of mass surveillance and especially its unlawful application may lead to the erosion of public trust in governments and state agencies (Westerlund et al. 2021).

3.2 Breakthroughs Needed

In the context of Digital Language Equality (DLE), the main challenges are linked to the inferior support and resources available for less common languages, and a need for improving the performance and capabilities of ST for these languages. The proliferation of ST, including areas with a high potential impact on individuals and large groups of users, also has to be considered in a wider context of policies governing ST and relevant fields and calls for major breakthroughs in terms of explainability for the critical methods and technologies. Policies and governance concerning the use of ST and data – in particular personal data – need to be kept up to date and on par with rapidly developing technologies and applications. In order to democratise

STs and to strengthen their position within LT and AI, the base of users should be widened. An increase in educational programmes, including in general AI, ML, NLP, and inter-disciplinary projects, is necessary for the continuous training of experts in these fields able to draw upon expertise in voice technologies but at the same time also in domain-specific fields, thus forming the links between them.

Training paradigms – For approaches requiring large amounts of annotated data, strategies and frameworks for joint (potentially distributed) data collection, improved annotation, and joint provision are needed. This not only concerns the collection but equally the storage and provision of such resources. A lack of commercial interest needs to be alleviated by public efforts to jump-start and boost efforts in low-resource languages to limit the threat of digital language extinction. From the perspective of data augmentation, the generation and use of synthetic data may provide a complementary alley in the creation or extension of datasets. Efficient use of transfer learning and fine-tuning, as well as work on algorithms and methodologies that use less data or provide more robust models with lower amounts of data, present promising alternatives to relieve the lack-of-data challenge. For specific fields of ST, improved use of unlabelled data in an unsupervised or semi-supervised manner (pre-training, self-supervised training) provides further possibilities (Lai et al. 2021). For several technologies, making better use of the hierarchical structure and relatedness of languages may be beneficial. Methods like one-shot learning or few-shot learning likewise provide promising approaches.

Access to and discoverability of training data – The need for large amounts of data severely limits the possibilities for small companies and niche players to compete and be able to develop their own solutions. A plethora of licensing agreements pose further obstacles to access datasets and resources. Simplification and harmonisation of these mechanisms would be highly beneficial. In the larger context of open data sharing and bringing digital technology to businesses, citizens and public administrations these issues connect with the EU's Digital Europe Programme.

Support for low-resourced languages – To provide first-rate ST in any language, additional high-quality datasets are essential. Creating a wide set may not be feasible in general, but could be achieved at least for several major European languages. New techniques for transfer learning and model adaptation from systems trained for resource-rich languages to systems able to function in languages with more reduced quantities of available data should enable the development of cutting-edge ST systems also for these languages. New architectures allowing the combination of resources from several languages in such a way that their commonalities are learned in a more robust way (by cross-lingual knowledge-sharing) and methods for the creation of multilingual or language-agnostic models which can be applied to a number of different languages are of utmost importance.

Confluence and context information integration – A tendency towards confluence – the combination of technologies and inclusion of a larger context – can be observed and also be assumed to play a more pronounced role in the future. The increased presence of conversational interfaces, a proliferation of chatbots combining ASR, NLP and TTS with an ever-increasing presence of AI in general, has modified not only the technical and commercial landscape but also the expectations of users, which have

been accelerated by increased time spent in home-office setups and virtual meetings. More powerful tools and greater capabilities also prompt the integration of upstream technologies such as summarisation or sentiment analysis with voice technologies. Speech synthesis is bound to become as emotional and persuasive as the human voice itself. Automatic translation may be used to bridge language barriers. Technologies will need to be integrated in a manner allowing for feedback loops and adaptation seamlessly. Models need to be dynamic and methods allowing for dynamic adaptation – learning and unlearning certain features – will need to be developed to account for flexible and continuously changing conditions. Areas of linguistics such as pragmatics or paralinguistics will need to be considered and integrated to a much higher extent to allow for more natural and human-like interaction. Adding emotions and affections into the recipes for HCI, recognising intent and taking into account a broad variety of contexts holds the potential to turn these interactions into truly human-like experiences. The components related to emotional understanding and empathy are especially relevant for systems functioning in social domains, such as healthcare, education, and customer service.

Explainability, transparency and privacy concerns – Trust in STs and in the use of data obtained by interacting with them may become a decisive factor in the adoption of technologies and success of individual market players. An increased interest in the transparency of data use and system functionality can be observed across the board in many areas of ML and AI. A fundamental question to be answered by providers will be where processing is performed and to what extent and purpose data is used to modify models. One end of the spectrum of processing is large, anonymous data-centres spread around the globe, the other is formed by strictly local processing on personal devices. On-premise solutions provided by companies or institutions form an intermediate setting. In all of these setups, the balance between capabilities and the requirements to achieve these capabilities will need to be determined and balanced against ethical concerns and personal and privacy-preserving arguments. The extent and amount of end-user control will be a crucial factor. Approaches like privacy-by-design accompanied by high ethical and legal standards may be determining factors in enabling trust, fostering adoption and leading to economic success.

Performance, robustness and evaluation paradigms – Driven by various national and international evaluations, standard performance measures have been defined on standard test sets. Current measures like the standard WER only take certain performance aspects into account and may need to be reconsidered, extended or complemented. Robustness and generalisability of ST components and models as well as standard evaluation sets for multiple languages and evaluation sets allowing the parallel evaluation of several technologies (all on the same dataset) should be devised. The topics of ageing and recency of data for evaluation sets need to be taken into consideration. In general, evaluation (as well as training) datasets should be viewed more as work in progress than static artefacts. Extension to further languages and language varieties, dialects and speaking conditions likewise should receive further attention to ensuring broad availability and adoption. Another needed innovation is a method for objectively measuring TTS results; such systems are currently assessed by means of subjective evaluations which are time-consuming and laborious.

Outreach – communities, non-experts – Recent years have witnessed an increase in interest in the democratisation of AI. The widespread application of ML and the well-known fact that experts in ML and AI have become scarce resources has led to the desire to empower a wider set of individuals to participate in the creation and use of these technologies. Toolkits and *do-it-yourself modelling* form part of the trend to democratise voice technologies. Approaches like Auto-ML aim to provide access to ML also for non-experts and align with strategies to allow a wider audience to participate in the process. As LTs are aggregated and applied to more complex settings, inter-disciplinary research and activities (for instance) from fields in the social sciences are becoming more relevant and synergies become apparent. Programmes and funding schemes to actively engage these communities and foster inter-disciplinary research would further boost developments.

Alignments with EU policies and policy breakthroughs needed – Copyright legislation is more restrictive in Europe than in other economic regions and countries, e. g., utilising closed captions from TV broadcasts or subtitles from a copyrighted film to train and evaluate ST models could enable access to high-quality language data if lawmakers could agree that training of models on copyrighted data constitutes fair use, as long as it does not diminish the value of the assets or reduce the profits reasonably expected by the owner. The pace of ST development in Europe could be further increased by introducing changes that enable the re-use of existing data, while at the same time ensuring that the value of the copyright owners is not impaired. GDPR introduced a new global standard that places an emphasis on individual rights and reflects European values, and as such contributes to building trust in AI. GDPR has had a *negative* impact on the majority of Europe's LT business and research activities (Smal et al. 2020). Furthermore, non-European AI firms have been able to operate free of GDPR constraints since then, giving them an economic advantage. One of the required breakthroughs relates thus to ensure that while individual rights are protected, the extent of these – in particular, in practical settings and day-to-day operations – does not go beyond the intended scope. Automatic, efficient and free anonymisation tools are required for all European languages.

3.3 Technology Visions and Development Goals

ST: the interface of the future – In many settings, voice provides the most natural way to interact with devices and appliances. The coming years will witness an increased advance in voice technologies to the point that interacting with automated systems will be virtually indistinguishable from communication with human beings in many cases. Interfaces predominately relying on typing, clicking and swiping will gradually transform into multimodal, or fully virtual interfaces including voice, shifting the task of adaptation from human users to computer systems. Compared to the other modalities currently dominating the HCI landscape, communication will encompass richer kinds of (linguistic and paralinguistic) information, including gender, age, emotional or cognitive state, health conditions or speaker-specific traits allow-

ing for more sophisticated and accurate speaker identification, modelling, adaptation and personalisation. These factors and their integration into HCI – as beneficial and powerful as they may be – also give rise to privacy and ethical concerns. They prompt questions of control, user understanding and intent when it comes to sharing information and the extent to which different kinds of information are transmitted and used in the future. Ensuing risks and the potential impact need to be carefully met and balanced with measures to increase security and trust through technical means as well as policy and legislative measures. Striking this balance will affect the adoption of a wide range of devices and services: from VAs in homes and phones, navigation and control systems in cars to cooperative office and work environments and systems supporting a wide range of business and leisure activities.

User and application contexts – A trend towards the integration of richer context is to be expected, regardless of the sub-field of voice processing. This concerns individual technologies and their combination. For TTS, to have a truly interactive experience when dealing with our devices, the integration of context will play a major role. To give just one example, the correct way to pronounce a message should be inferred from the context or the previous steps of a dialogue. Technologies will need to be sensitive to the user's character, state, mood and needs and adapt themselves accordingly. Potentially, they will also need to take into account other participants' states in case of group activities such as business meetings. Topics of pragmatics will be reflected by all technologies. Rather than individual communication turns, complete conversations with history and context will be the norm.

Addressing existing technological gaps – Continued efforts towards better understanding and modelling human speech perception might result in sophisticated ASR addressing several of the limitations and gaps identified in current approaches. Improved handling of audio conditions currently perceived as difficult (e. g., multiple simultaneous speakers in noisy environments speaking spontaneously and highly emotionally in a mix of languages) will be possible thanks to such advances. A wider deployment and further popularisation of ST will require solutions that offer high robustness, low latency, efficient customisation and the ability to provide possible equal support for a diverse set of speakers.

ST integration – An intimate relation of ASR, SID and TTS with downstream Natural Language Understanding (NLU) technologies is needed to allow the correct interpretation of the input. A combination of technologies to interact in multimodal ways (including visuals) and the efficient combination of inter-linked models will be able to guarantee the best experience possible. The successful combination will result in an enhanced easiness and naturalness of use, hiding individual components and allowing systems to be perceived as assistants using natural language much in the way that human assistants would.

Multimodal models – Recently introduced NN architectures support encoding and decoding schemes of various modalities, e. g., Perceiver IO (Jaegle et al. 2021). Despite being task-agnostic, the model provides competitive results on modalities such as language, vision, multimodal data, and point clouds. In the near future, this type of architecture is expected to be used in a range of applications where multimodal content needs to be jointly analysed. Furthermore, a future line of work that can eas-

ily be envisaged is the training of a single, shared NN encoder on several modalities at the same time, and only using modality-specific pre- and post-processors.

Development pace – The pace of development in voice-based technologies is driven by general advances in ML and associated hardware as well as domain-specific advances in speech perception and production. The former can be expected to accelerate even more due to general interest in ML and AI from a wide portfolio of domains. Advances in transfer learning, reinforcement learning, fine-tuning, the use of pre-trained models and components as well as the arrival of platforms such as Hugging Face have created additional momentum. The extension of GPU capabilities can likewise be expected to continue at a fast pace.

Training and evaluation – Further improvements introduced in the process of creation and distribution of ever-growing, ever more coherent and diverse datasets can be expected. These will include large, multilingual, multi-domain and multimodal datasets, which will become de facto standard sets for training and evaluation. We will witness an increase in labelling efficiency, a wider adaptation of continuous learning, self-adaptation and self-modification paradigms. While datasets will continue to grow, the quality and amount of data of high- versus low-resourced languages are unlikely to converge in the short term. The development of more complex and multifaceted datasets calls for more comprehensive evaluation and quality criteria: a shift that would change the focus from an individual technology to an end-user assessment of an experience while conducting a specific task in a non-laboratory environment and within a specific operational and personalised contexts.

Infrastructure, hardware – Extrapolating from the current trends a further rapid increase in the capacities of ST-related hardware and infrastructure can be foreseen (faster communication networks, higher bandwidths). Further popularisation of ST solutions in the context of the Internet of Things (IoT), and a new set of voice-enabled devices will be available to users at work, leisure and commercial settings. These developments create additional challenges related to load and scalability of the underlying infrastructure, hardware and networks. Moving computation to edge devices will also continue to be a trend in the near future.

Privacy, accountability and regulations – The future development of ST and the wider LT field will be strongly influenced by the regulations governing the collection, storage, transmission, and use of personal data. In the context of European AI companies and research institutes, the pace of development appears to be particularly influenced by current regulation schemes. Lawmakers' decisions will thus have to consider the wide and profound impact of their regulations: on the protection of citizens' personal data and privacy on the one hand, and on the wider field of AI technologies and the comparative advantages and disadvantages vis-à-vis other geopolitical regions on the other. Extrapolating from current regulations concerning user privacy, and differences in data collection and use, it seems probable that the divide between the EU and non-EU countries will continue to grow. It is unlikely that a consensus or standardisation between competing regions will be found. With the growing presence of ST and AI in general, increased concerns about hidden flaws, shortcomings and baked-in biases of such systems are gaining momentum. Whereas citizens and academia may work towards enhancing transparency and mechanisms

that may be able to avoid certain phenomena, the industry may work towards obfuscation and hindrance of these mechanisms. A sequence of scandals and growing interest in issues of ethics and privacy have led to an increased awareness in society of this issue. Trust in technology is a key ingredient for the adoption of technologies by a large portion of the population. Transparency in how privacy is integrated into technologies is a crucial ingredient to earning trust. Privacy-by-design beyond mere statements may become a decisive factor for technology uptake and market success.

Disclosure of the use of AI/ST – Due to the ever more human-like nature of ST, the use of AI technologies should be disclosed at the earliest stage possible for all transactions and applications. Making users aware of what they interact with can be regarded as a fundamental step in the creation of more transparency. This will not prevent humans from attributing personhood to machines or hinder human-like communication, but present an ethical and transparent frame around such settings.

Audits of algorithms and models – Auditors will have to be independent for this to make sense and not open the door to even more secretive and evasive behaviour by companies. Federal agencies or boards may be required to preside over such activities. Standard test sets and tests may have to be created and applied.

Impact assessments of the introduction of such technologies – The concept of measuring impact and potential harm is firmly established in fields such as the environment. Similarly, algorithmic impact assessments need to cover a broad range of factors, with ST and NLP focusing on language- and language use-related aspects.

Public repositories of incidents where AI/NLP caused harm – Public repositories and ways to report problematic uses of AI would allow the identification of repeat offenders and act in case of recurring problems. Furthermore, making such cases known publicly may serve as an incentive to correct or prevent them.

Effects on society, workplace – The discussion about which jobs or areas within domains are likely candidates to be replaced by AI carries over to the domain of speech processing – as well as to NLP in general – as they form a core element of AI. Issues concerning automation and job replacement and the ensuing policy-making and social ramifications thus also directly concern ST and their perception.

Pervasiveness – A further spread and ubiquitous presence of voice-based technologies, and wider deployment of ST across a multitude of services and devices due to a reduction in size and integration into wearable and virtual environments can be expected. This may also concern further persons being in the vicinity of such deployments who may be involved indirectly by someone else's use of ST.

Future applications – ST in combination with other NLP and AI technologies will pave the way for intelligent applications with human-like capabilities and the potential for disruptive innovation in various sectors. Intelligent assistants and chatbots currently provide the leading paths towards general and broad adoption. Future applications will be expected to understand a user's intents over sequences of interactions, completely eliminating perceived boundaries between individual technologies. STs are already being used by multiple industries to increase self-service functionalities, reduce average handling time, increase availability and reduce employee costs.

Personalised Voices – Voices for TTS will be generated for any language and be fully customisable. In the same way as we can now personalise avatars in video

games, we will be able to set every aspect of the synthetic voice to suit the characteristics we prefer for each situation. Moreover, TTS technology will extend, and speech will be generated not only from text but also from other input information that could be more convenient for some users who do not have easy access to text or for some situations (e. g., requiring privacy). Multi-modal systems will allow the generation of speech from lip-reading, articulatory data acquired by diverse technologies such as electromyography, permanent magnet articulography and other silent speech interfaces, and even cerebral activity with brain-computer interfaces.

Ambient intelligence – Viewing ST as a means for intelligent interaction, integrating nuanced and fine-grained context and input from multiple modalities can be expected to lead to more human-like systems where the perception of individual components will blur into an overall experience for end-users. Such combinations may be a step towards a broader kind of AI as opposed to the narrow, highly-specialised versions in use today.

3.4 Towards Deep Natural Language Understanding

In many instances, the most natural manner for humans to interact with machines is through voice, for issuing commands or queries as well as generating responses and statements. Certain types of scenarios (e. g., limiting the interaction to small, handheld devices) may call for voice-only interaction, whereas others (e. g., allowing for feedback via large screens, augmented- or virtual-reality environments) may favour multimedia settings, permitting the flow of information across different modalities in parallel. Other scenarios may ask for communication completely without the use of audio, in particular when considering special needs and inclusive communication.

STs play a role in the ingestion of information, by acting as a kind of sensor conveying linguistic as well as paralinguistic inputs and converting them into structured information. Equally, their use concerns the output of information in auditive form (speech, but also non-speech, e. g., confirmations) to communicate with human users. Both directions of the flow of information apply to HCI as well as H2H interaction in the case of groups of human users interacting with each other or with computers, e. g., during meetings with intelligent assistants for transcription, translation and summarisation. STs thus form an intermediate interface layer between humans and machines. Inbound (auditive) information is captured and enriched by ST before being passed on to downstream NLU processing. Outbound information is enriched, transformed and eventually realised as audio based on content, structure and metadata provided by semantic components. The semantics and interpretation of utterances as well as the generation of appropriate responses based on a logical representation and state of a conversation fully reside within the scope and components of NLU and technologies such as dialogue managers (to carry out conversations) or knowledge graphs (networks for semantic representations). As such, STs provide essential contributions to the functioning of NLU in the input and output directions but they do not perform any semantic processing (understanding) themselves.

Visual cues such as gestures or manual articulation (sign language) may replace the audio-element of ST when operating in noisy environments or involving hearing-impaired or deaf people. Visual processing technologies assume the roles of ST in these cases. The combination of modalities is also possible and may be appropriate or imperative depending on the actual context, such as working environments requiring a hands-free operation. The contribution of ST towards achieving deep NLU may thus lie in the improvement and extension of the individual technologies (both from accuracy as well as a language- and domain-coverage perspective), their integration into E2E systems allowing for joint operation and optimisation, including different kinds of knowledge sources and their flexible and dynamic configuration depending on the state and context of an application or user. Approaches including the combination of several modalities for input and output may likewise prove beneficial in the context of achieving deep NLU. In many cases, the real power of NLU will become clear when it is part of a complex system functioning as a human-like counterpart in communication: exhibiting context, history and elements of general intelligence. However, it may also come about that NLU is overshadowed by the cognitive downstream processing and eventually perceived as a mere commodity. The element of admiration and awe on the part of the user will then concern the complete system performance, with NLU itself disappearing in importance as a small part of a much larger and more complex intelligent system.

4 Summary and Conclusions

The substantial advances made in the field of STs over the past decades hold the potential for disruptive innovation in many areas and application domains. Combined with the progress of related fields, they provide the basis for the broad adoption of speech and voice as the primary modality for interacting with computer systems as part of larger and more complex systems modelling human-like communication and interaction. This chapter outlined several research fields and business domains that provide promising areas for the use of ST and their inclusion into larger solutions yielding more natural means of communication. Several issues and challenges have been identified which need to be resolved to make this promise materialise. Below we summarise the key elements identified and provide recommendations for possible future actions. All these strands of progress can aid in supporting the overarching goal of achieving DLE in Europe by providing services made possible by these technologies to larger multilingual audiences at similar levels of scope and performance.

Training data is still a key factor as long as supervised paradigms prevail. Accessibility is often limited, or even locked, with individual actors amassing massive amounts of data, effectively creating monopolies for certain markets. Licences and regulation as well as interoperability and compatibility of data resources and providers remain obstacles that need to be overcome. Methods not relying on vast amounts of data are an active area of research.

Even though the range of languages supported by ST has increased dramatically over the past decades, English still holds a privileged position. The creation of resources for further languages and dialects (some may only be spoken) is ongoing; the investigation of phenomena that are only present in other language families is also an active area of research. The creation of multilingual or language-agnostic models provides further avenues for improvement.

A trend of integrated E2E models into one combined overall model can be observed. Training takes place in a single framework rather than individually, capitalising on joint factors. Considerable progress in performance has been made through this approach which can be expected to continue. The integration of semantic components such as NLU or knowledge graphs into these frameworks may provide additional elements required for intelligent interaction.

In current applications, different components operate in an independent and isolated manner. The dynamic inclusion and integration of context would allow STs to operate on a significantly higher level of accuracy, eliminating errors and narrowing down alternatives. Various ways for the fusion of information have been investigated but have not effectively come to fruition. Parallel systems for multiparty conversations and multimodal approaches may provide ways forward.

STs primarily address the voice modality for interacting with computers. Combining STs with multimodal inputs and outputs may provide a basis for next-generation HCI. The inclusion of gestures, facial expression, emotions or haptics, and the generation of multimodal outputs reflecting these elements may result in a richer and more natural user experience and lead to wider adoption and acceptance of ST.

Although established measures allow quantification of progress in ST, they may only tell part of the story when it comes to real-world applications and the combination with downstream processing. In many fields of ST, performance has reached (near-)human levels under controlled conditions with progress being significant in theory but often only marginal when translated into reality. A shift towards increasing robustness and generality of results may prove beneficial at this stage.

Recent progress and an abundance of ST in chatbots may evoke expectations of ST being a mere commodity and raise unrealistic expectations on the part of users. STs perform considerably worse when applied to conditions unlike those for which they were originally created. Accordingly, adaptation and customisation to special domains provide opportunities for specialists. Expectation management and open communication about the possibilities but also limitations from the ST community may help set expectations to realistic and practical levels.

The interest and concern about fairness and biases of models and ethical issues relating to their use have been receiving increased attention. Methods for detecting biases and de-biasing need to be improved and are expected to become a more active area of development. Furthermore, access to ST for people with disabilities and impairments needs to be extended. Triggered by an increased interest in the fairness of AI systems (e. g., assessments of job applications, prison-parole, credits), applications continue to be subjected to scrutiny. Users demand explanations on the capabilities and functioning of ST. Results are questioned with some application areas demanding audits of models and algorithms. Technical issues need to be

addressed and accompanied on the policy-making and legislative levels. Standardisation of evaluations and publication of results may function as motivating factors for providers to address these issues more thoroughly.

With the current and near-future state of ST, many businesses, political parties and ideological movements may develop conversational agents as a ubiquitous representatives to convey their agenda and sway public opinion to get support for their cause. Situations where the agents' identity is known or hidden should be clearly distinguished. Cases where a company or party is represented by a single conversational agent, or by hundreds or even thousands to create a representation of mass support, should be marked. Scandals, data leaks and an increase in cyber-crime have brought issues of security and privacy to the fore. Devices are ever more pervasive, taking ST into people's offices and homes. IoT and wearables further accelerate this trend. Users are becoming increasingly wary of the risks and undesired effects related to the introduction of ST. Clandestine ways of data collection and eavesdropping infringing privacy are rightly exposed and castigated by the media. Actors risk suffering dire consequences if they do not respond and put corrective measures into place. The balance between convenience and privacy will remain a fluid one to be negotiated repeatedly and on multiple levels.

The legislation governing the acquisition, storage, transmission, and use of personal data has a significant impact on the future of ST and the wider LT area. Extrapolating from current trends, the gap between the regulations used in different regions will continue to widen. As AI technologies play a critical role in creating competitive advantages across a wide range of human activities, it is unlikely that competing countries and regions will be able to reach a broad, far-reaching agreement, resulting in one standardised set of regulations. Lawmakers' decisions will thus have to consider a wide and profound impact of their regulations, on the protection of citizens' personal data and privacy on the one hand, and on the pace of development in the broader field of AI technologies on the other: research, development and application and the comparative advantages and disadvantages vis-à-vis other regions and global centres of AI technology development.

As technologies need to be accepted by society in order to be adopted, advancements as described in this chapter are not exclusively technical ones, but need to be accompanied by progress from the humanities. Multi-disciplinary approaches, as demonstrated by the rise of the digital humanities, may prove advantageous also in these scenarios. As systems become natural companions, the fields of psychology, neuroscience and philosophy bring new aspects and visions to the agenda and inspire novel approaches. Fear and anxieties generated by overly aggressive marketing, science-fiction and disinformation need to be met with prudent transparency, adequate management of expectations and accompanying policy measures. An inclusive approach akin to making ST (and AI) visible, transparent and understandable to a larger public – a kind of AI-literacy in the sense of media-literacy – may be a strong supporting topic for all the above-mentioned domains. People have always tended to humanise machines. Powerful systems formed by the combination and integration of technologies and components described above may effectively be attributed human-like qualities and personhood by their users. Ethical aspects of such interaction must

be addressed in parallel with technological progress. Transparency (e. g., chatbots introducing themselves as machines) and openness are among the key factors to be considered when leaving users a freedom of choice rather than imposing technology on them. This certainly reaches far beyond ST but rather concerns AI in general.

References

Backfried, Gerhard, Marcin Skowron, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernaez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, and Petr Schwarz (2022). *Deliverable D2.14 Technology Deep Dive – Speech Technologies*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/speech-deep-dive.pdf.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli (2020). "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *NIPS'20: Proc. of the 34th Int. Conf. on Neural Information Processing Systems*, pp. 12449–12460.

Bommasani, Rishi et al. (2021). *On the Opportunities and Risks of Foundation Models*. arXiv: 2108.07258 [cs.LG]. https://arxiv.org/abs/2108.07258.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Draxler, Christoph, Henk van den Heuvel, Arjan van Hessen, Silvia Calamai, and Louise Corti (2020). "A CLARIN Transcription Portal for Interview Data". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3353–3359. %7Bhttps://aclanthology.org/2020.lrec-1.411%7D.

Garrido, Miguelángel Verde (2021). "Why a Militantly Democratic Lack of Trust in State Surveillance can Enable Better and More Democratic Security". In: *Trust and Transparency in an Age of Surveillance*. Routledge, pp. 221–240.

Jaegle, Andrew, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira (2021). "Perceiver io: A General Architecture for Structured Inputs & Outputs". In: *arXiv preprint arXiv:2107.14795*.

Jelinek, Frederick (1998). *Statistical Methods for Speech Recognition*. Cambridge: MIT Press.

Lai, Cheng-I Jeff, Yang Zhang, Alexander H Liu, Shiyu Chang, Yi-Lun Liao, Yung-Sung Chuang, Kaizhi Qian, Sameer Khurana, David Cox, and Jim Glass (2021). "PARP: Prune, Adjust and Re-Prune for Self-Supervised Speech Recognition". In: *Advances in Neural Information Processing Systems* 34, pp. 21256–21272.

Pessanha, Francisca and Almila Akdag Salah (2022). "A Computational Look at Oral History Archives". In: *Journal on Computing and Cultural Heritage* 15.1.

Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu (2021). "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Ruiz, Nicholas, Mattia Antonino Di Gangi, Nicola Bertoldi, and Marcello Federico (2019). "Assessing the Tolerance of Neural Machine Translation Systems Against Speech Recognition Errors". In: *CoRR* abs/1904.10997. arXiv: 1904.10997. http://arxiv.org/abs/1904.10997.

Smal, Lilli, Andrea Lösch, Josef van Genabith, Maria Giagkou, Thierry Declerck, and Stephan Busemann (2020). "Language Data Sharing in European Public Services – Overcoming Obstacles and Creating Sustainable Data Sharing Infrastructures". In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association, pp. 3443–3448. https://aclanthology.org/2020.lrec-1.422/.

Stahl, Titus (2016). "Indiscriminate Mass Surveillance and the Public Sphere". In: *Ethics and Information Technology* 18.1, pp. 33–39.

Tang, Dengke, Junlin Zeng, and Ming Li (2018). "An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals". In: *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*. ISCA, pp. 162–166. https://doi.org/10.21437/Interspeech.2018-2581.

Tomanek, Katrin, Françoise Beaufays, Julie Cattiau, Angad Chandorkar, and Khe Chai Sim (2021). "On-Device Personalization of Automatic Speech Recognition Models for Disordered Speech". In: *arXiv preprint arXiv:2106.10259*.

Valle, Rafael, Jason Li, Ryan Prenger, and Bryan Catanzaro (2020). "Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens". In: *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, pp. 6189–6193.

Valle, Rafael, Kevin J. Shih, Ryan Prenger, and Bryan Catanzaro (2021). "Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Wang, Yuxuan, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous (2018). "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research, pp. 5167–5176. http://proceedings.mlr.press/v80/wang18h.html.

Westerlund, Mika, Diane A Isabelle, and Seppo Leminen (2021). "The Acceptance of Digital Surveillance in an Age of Big Data". In: *Technology Innovation Management Review* 11.3.